PAPER

CENSREC-3: An Evaluation Framework for Japanese Speech Recognition in Real Car-Driving Environments

Masakiyo FUJIMOTO[†*a)], **Kazuya TAKEDA**[††b)], ***and*** **Satoshi NAKAMURA**[†**c)], ***Members***

SUMMARY This paper introduces a common database, an evaluation framework, and its baseline recognition results for in-car speech recognition, CENSREC-3, as an outcome of the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group. CENSREC-3, which is a sequel to AURORA-2J, has been designed as the evaluation framework of isolated word recognition in real car-driving environments. Speech data were collected using two microphones, a close-talking microphone and a hands-free microphone, under 16 carefully controlled driving conditions, i.e., combinations of three car speeds and six car conditions. CENSREC-3 provides six evaluation environments designed using speech data collected in these conditions.

key words: *noisy speech recognition, common evaluation framework, in-car speech database, CENSREC-3*

1. Introduction

Recently, progress in speech recognition technology has been brought about by the advent of statistical approaches and large-scale corpora. Furthermore, it is also widely known that progress has been accelerated by the U.S. DARPA projects [1] initiated in the late '80s. This involves project participants competitively developing speech recognition systems for the same task, using the same training and test corpus.

However, current speech recognition performance must still be improved if the system is to be exposed to noisy environments, where speech recognition applications might be used in practice. Therefore, noise robustness is an emerging and crucial factor to be solved for speech recognition techniques.

With regard to the noise robustness problem, there have been two major evaluation projects, SPINE1, 2 [2] and AURORA [3]–[9]. The SPINE (SPeech recognition In Noisy Environments) project was organized by the U.S.'s DARPA, with SPINE1 in 2000 and SPINE2 in 2001. The task included spontaneous English dialog between an operator and a soldier in a noisy field to evaluate spontaneous continuous speech recognition in noisy environments. The results of the project brought many improvements to continuous noisy speech recognition, though that task seems quite specialized and a little difficult to handle.

On the other hand, the European Telecommunications Standards Institute (ETSI) AURORA group initiated a special session in the EUROSPEECH conference. They are actively working to develop standard technologies under ETSI for distributed speech recognition [10]. In parallel with their standardization activities, they have distributed to academic researchers a noisy connected speech corpus based on TI-digits [11] with baseline HTK (HMM Took Kit) [12] scripts for further noisy speech recognition research. To date, AURORA2 [3] (a connected digit corpus with additive noise), AURORA3 [4]–[7] (an in-car noisy digit corpus), and AURORA4 [8], [9] (a large-vocabulary continuous-speech recognition corpus with additive noise (noisy Wall Street Journal, vocabulary size: 5,000)) have been distributed with HTK scripts, which can be used to obtain baseline performance and even improvements over the baseline results [13].

The authors voluntarily organized a special working group in October 2001 under the auspices of the Information Processing Society of Japan in order to assess speech recognition technology in noisy environments. The focus of the working group included the planning of comprehensive fundamental assessments of noisy speech recognition, standardized corpus collection, evaluation strategy developments, and distribution of standardized processing modules. As an outcome of the working group, we have already been produced the Japanese AURORA-2, AURORA-2J [14], which comprises the English digits translated into Japanese.

This paper introduces a common database, an evaluation framework, and its baseline recognition results for in-car speech recognition, CENSREC-3 (Corpus and Environments for Noisy Speech RECognition), as a sequel to AURORA-2J[***].

AURORA-2J (CENSREC-1) was designed as a common evaluation framework for noisy speech recognition. However, the noise environments provided by AURORA-2J were simulated environments; namely, speech and noise signals were recorded independently and the noisy speech

Manuscript received January 18, 2006.
Manuscript revised June 5, 2006.

[†]The authors are with the ATR Spoken Language Communication Research Laboratories, "Keihanna Science City", Kyoto-fu, 619-0288 Japan.
[††]The author is with Nagoya University, Nagoya-shi, 464-8603 Japan.
[*]Presently, with NTT Communication Science Laboratories.
[**]Presently, with National Institute of Information and Communications Technology.
a) E-mail: masakiyo@cslab.kecl.ntt.co.jp
b) E-mail: takeda@is.nagoya-u.ac.jp
c) E-mail: satoshi.nakamura@atr.jp
DOI: 10.1093/ietisy/e89–d.11.2783

[***]AURORA-2J is regarded as a part of the CENSREC series and has been given an alternative name, CENSREC-1.

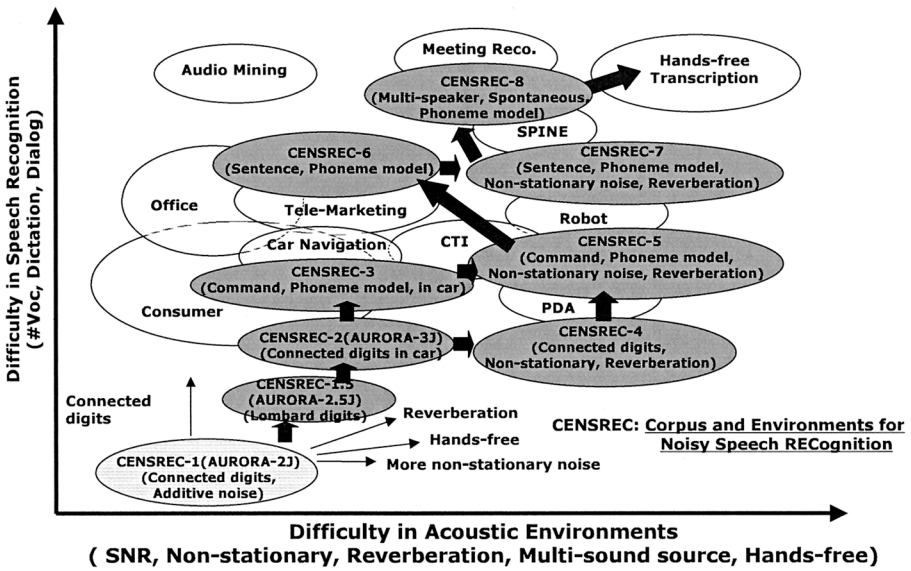


Fig. 1 Roadmap of the CENSREC database. The gray ellipses indicate evolution of the CENSREC database. The white ellipses indicate target applications or tasks for each CENSREC database.

data were artificially generated by adding the noise to the speech. Such simulated noisy speech data are widely used for evaluating noisy speech recognition due to the eases of constructing the evaluation environment. However, the transformation of speaking style is not considered and is neglected in the simulated environments. Usually, a speaker speaks loudly or shouts when he is in a noisy environment. In this case, the transformations of energy dynamic range, frequency characteristics, speaking speed, and so on ought occur in uttered speech signals. Thus, to involve these transformations of speaking style, it is necessary to collect the speech data in real noise environments.

The speech data of CENSREC-3 were collected in real car-driving environments to evaluate in-car speech recognition. Recently, in-car speech recognition, which is one of the candidates for a hands-free interface for controlling electric devices in a vehicle, has attracted attention as an extremely important technique from the viewpoints of safety and convenience. An in-car environment is a typical noise condition of speech recognition, and transformations of speaking style are also observed. Thus, careful design of an evaluation framework for in-car speech recognition is an important concern.

Incidentally, as a database for evaluating in-car speech recognition, we have designed not only CENSREC-3 but also CENSREC-2 [15]. Although the speech data of both CENSREC-2 and CENSREC-3 are collected in real car-driving environments, the purpose of each one's database design is different. Figure 1 shows our defined roadmap of the CENSREC database. In the figure, the position of CENSREC-2 is lower than CENSREC-3 in the layer of difficulty of speech recognition (the measure for the vertical axis). CENSREC-2 was designed as the evaluation framework for connected digit recognition in real car-driving environments. However, digit recognition is not widely applicable to tasks necessary for driving in the real world. Thus, the target application of CENSREC-2 is not assigned to the roadmap due to its inapplicability. On the other hand, CENSREC-3 has been designed specifically as the evaluation framework for isolated word recognition in real car-driving environments. Since the main target application of CENSREC-3 is human-voice (hands-free) control of car navigation systems, CENSREC-3 is an evaluation framework that assumes speech-oriented man-machine communication in a range of different car environments.

Speech data of the CENSREC-3 were collected using two microphones, a close-talking microphone and a hands-free microphone, under 16 carefully controlled driving conditions, i.e., combinations of three car speeds and six car conditions. CENSREC-3 provides six evaluation environments designed using speech data collected in these conditions. Finally, this paper shows the evaluation results for CENSREC-3 by using ETSI standard DSR front-end ES 202 050 [16], i.e., advanced front-end. We also analyze the crucial environments for practical use of in-car speech recognition through the evaluations of CENSREC-3.

2. Data Recording

The CENSREC-3 database is composed of part of the database collected by the Center for Integrated Acoustic Information Research (CIAIR) [17].

2.1 Vocabulary

The speech recognition task of the CENSREC-3 database is isolated word recognition in real car-driving environments. Table 1 shows a list of 50 words recorded for test data, which are classified into several groups, e.g., control commands, song or musician names, street or highway names, place

Table 1 A list of 50 recorded words.

digital_locker	ninsho_kaishi
2001/1/1	yamada_tarou
kensaku_shuryo	ansho_bango
0123	4567
8901	2345
6789	contents
eiga	Hitsuji_tachino_chinmoku
Sound_of_music	game
Pack_man	ongaku
jpop	konsyu_no_top10
genre_betsu_kensaku	pops
rock	Beatles
senkyoku	Yesterday
Let_it_be	haishin_kaishi
ferry_annai	jikoku_hyo
dai2bin_wo_yoyaku	net_news
topics	onsei_yomiage
tenki_yohou	koutsu_jouhou
Kanagawa_ken	Yokohama_shi
Naka_ku	Toukyou_to
Setagaya_ku	Syuto_kousoku
Touhoku_jidoushadou	Seven_eleven
Uniqlo	Star_bucks
hotel_ichiran	Pacific_hotel
yoyaku_hyo	service_shuryo

names, shop names, and so on. In each group, we selected some frequently used words.

2.2 Speech Data Recording

In-car speech data were collected in a vehicle specially equipped with seven microphones mounted as shown in Fig. 2. Microphones 1 (driver) and 2 (navigator) were close-talking headset microphones, microphones 3 and 4 were attached to the dashboard, and microphones 5, 6, and 7 were fixed to the ceiling of the vehicle. The speech data recorded with the driver's close-talking (CT) microphone (no.1: SONY ECM77B mounted on SENNHEISER HMD410) and the hands-free (HF) microphone attached to the ceiling of the driver's seat (no.6: SONY ECM77B) are used for CENSREC-3 [17].

Microphone 6 is the closest hands-free microphone to the driver (speaker). Thus, it can capture the speech signal with higher quality, e.g., higher SNR and lower distortion or reverberation, than other hands-free microphones. Based on this information, it is reasonable to use the closet hands-free microphone for in-car speech recognition. In addition, the microphone attached to the ceiling is also used as the hands-free microphone for AURORA3. From the above advantage and with respect to AURORA3, we chose microphone 6 as the hands-free microphone to be used in this work.

The recording conditions for the evaluation data are shown in Table 2. Speech data were recorded under 16 environmental conditions using combinations of three different vehicle speeds (idling, low-speed driving on a city street, and high-speed driving on an expressway) and six in-car environments (normal, with hazard lights on, with the air-conditioner on (fan low/high), with the audio CD player

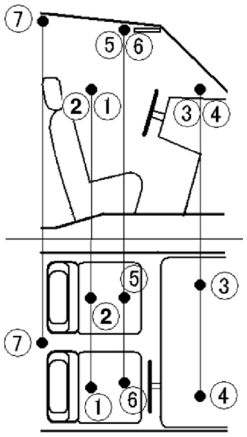


Fig. 2 Microphone positions for data collection: Side view (top) and top view (bottom).

Table 2 Recording environments for test data.

Car speed	In-car conditions
Idling (quiet)	Normal, Hazard lights on, Fan (low), Fan (high), Audio on, Windows open
Low speed	Normal, Fan (low), Fan (high), Audio on, Windows open
High speed	Normal, Fan (low), Fan (high), Audio on, Windows open

on, and with windows open). In these conditions, the "Hazard lights on" condition is used only when idling. Thus, we recorded the speech data under the six car conditions for idling and five car conditions for low- and high-speed driving ($6 + 5 \times 2 = 16$ conditions). A total of 14,216 utterances spoken by 18 speakers (8 males and 10 females) were recorded by each microphone.

For training, drivers' speech of phonetically-balanced sentences was recorded under two conditions: while idling and while driving on a city street with a normal in-car environment [17]. A total of 14,050 utterances spoken by 293 drivers (202 males and 91 females) were recorded by each microphone. The number of sentences per driver was 50 (idling) or 25 (driving). The drivers uttered the sentences by reading the written texts while idling. In the case of recording while driving, the sentences were divided into some short segments to be easily memorized by the drivers. The drivers uttered each segment of the sentences after listening to the recorded instruction speech played via headphones. Speech data of the segments were saved in separate files. In the CENSREC-3 database, since an "utterance" is defined as the speech data saved in one file, the number of utterances while driving is larger than in the case of idling even though the actual amount of recorded data (the length of recorded data) is less than in the case of idling.

Collecting speech comprising phonetically-balanced sentences takes longer time that of just words. In collecting the phonetically-balanced speech while driving, the driver's concentration may deteriorate, so for safety reasons we decreased the amount of data collection per driver while driving. In addition, in the case of high-speed driving, the driver

Table 3 Average SNR in each environment (dB).

Data	Training data		Test data											
Condition	Normal		Normal		Fan (low)		Fan (high)		Audio on		Windows open		Hazard lights on	
Microphone	CT	HF	CT	HF	CT	HF	CT	HF	CT	HF	CT	HF	CT	HF
Idling	40.52	18.20	41.19	16.75	32.86	11.01	25.76	5.47	31.46	11.57	29.92	8.63	33.50	12.48
Low speed	36.21	11.25	38.39	10.96	32.11	8.67	22.64	2.75	30.16	10.20	23.35	3.92	—	—
High speed	—	—	30.11	5.89	28.58	3.59	21.65	1.46	24.46	5.08	21.28	0.91	—	—

needs to concentrate more than when driving slowly, so we did not collect speech comprising sentences while driving at high speed. The speech signals for training and evaluation were both sampled at 16 kHz, quantized into 16-bit integers, and saved in the little-endian format.

Table 3 shows the average SNR (Signal to Noise Ratio) in each recording condition. In the table, we can see that the average SNR of speech data recorded by the close-talking microphone is high. The SNR is higher than 20 dB even when the recording condition is high-speed driving with the windows open, which is the worst recording condition with the lowest SNR. On the other hand, the average SNR of speech data recorded by the hands-free microphone is low. The SNR is less than 20 dB in all the recording conditions. In addition to the hands-free case, the worst recording condition is high-speed driving with the windows open, for which the SNR is approximately 0 dB.

2.3 Analysis of Observed Noise

From Table 3, we can see that the influence of noise was negligible for the speech data recorded by the close-talking microphone because SNRs of the data are considerably high. On the other hand, the speech data recorded by the hands-free microphone were seriously degraded by adverse noises. The adverse noises are classified into two types: driving (engine) noise and several ambient noises. In the speech data recorded by the hands-free microphone, we analyzed the adverse noise characteristics in each in-car condition as follows:

Normal: The driving noise has a typical car noise characteristic, namely the high energy distribution around the low-frequency components as shown in Fig. 3 (a). From the figure, we can see that the low-frequency energy increases in connection with car speed. The characteristic of driving noise with time is almost stationary; thus, the "Normal" condition at each driving speed can be regarded as a stationary noise environment.

Hazard lights on: Impulsive noise caused by the blinking of hazard lights is overlapped on the data (speech and engine noise) approximately every 350 msec. The duration of impulsive noise is less than 20 msec.

Fan low / high: Two levels of blowing noises by the air-conditioner overlap the data. Since in the "Fan high" condition the air-conditioner blows more strongly than in the "Fan low" condition, the SNR of the "Fan high" condition is lower than that of "Fan low" as shown in Table 3. The blowing noises have high energy not only in the low-frequency component but also in the high-frequency component as shown in Fig. 3 (c) and (d), from which it is clear that the spectra have similar shapes at each driving speed. Especially, the shapes for the "Fan high" condition are almost the same, meaning that the speech data recorded in the blowing air-conditioner conditions are strongly affected by the blowing noise. In addition, the characteristics of the blowing noise with respect to time are almost stationary. Consequently, these conditions at each driving speed can be regarded as stationary noise environments.

Audio on: Music played by the CD player overlaps the data. Samples of average log-power spectra are shown in Fig. 3 (e). From that figure, the spectra have different shapes at each driving speed. Here, frequency structures of the overlapping song usually have time-varying characteristics. In addition, these time-varying characteristics depend on the song, meaning that frequency structures dynamically change according to the overlapping song and the time slices.

Windows open: Noise of cutting through the wind is observed. This noise has high energy not only in the low-frequency components but also in the high-frequency ones, just as for the blowing noises. From Fig. 3 (f), we can see that the energy of the noise is strongly affected by car speed. The noise of cutting through the wind has almost stationary characteristics with respect to time.

In each environment, the noises produced by oncoming vehicles sometimes overlap on the data. In this case, the noise exhibits non-stationary characteristics.

3. Design of the Evaluational Framework

CENSREC-3 provides six evaluation environments for speech recognition using the speech data collected in the various recording conditions described in the previous section[†]. The six evaluation environments, called "condition 1 . . . 6," are defined based on the levels of acoustical mismatches between training data and testing data (the differences of driving speeds or/and microphones). Each evaluation environment consists of the acoustical conditions marked by a circle (◯) in Tables 4 and 5. In those tables, the speech data are simply divided according to car speed

[†]Note that a license fee is required ONLY FOR part of the training data, which were collected by using a HANDS-FREE MICROPHONE. You should pay the license fee if you wish to use a part of the charged data collected by using distant-talking microphones, although the CENSREC-3 DVD disk includes both free and charged speech data.

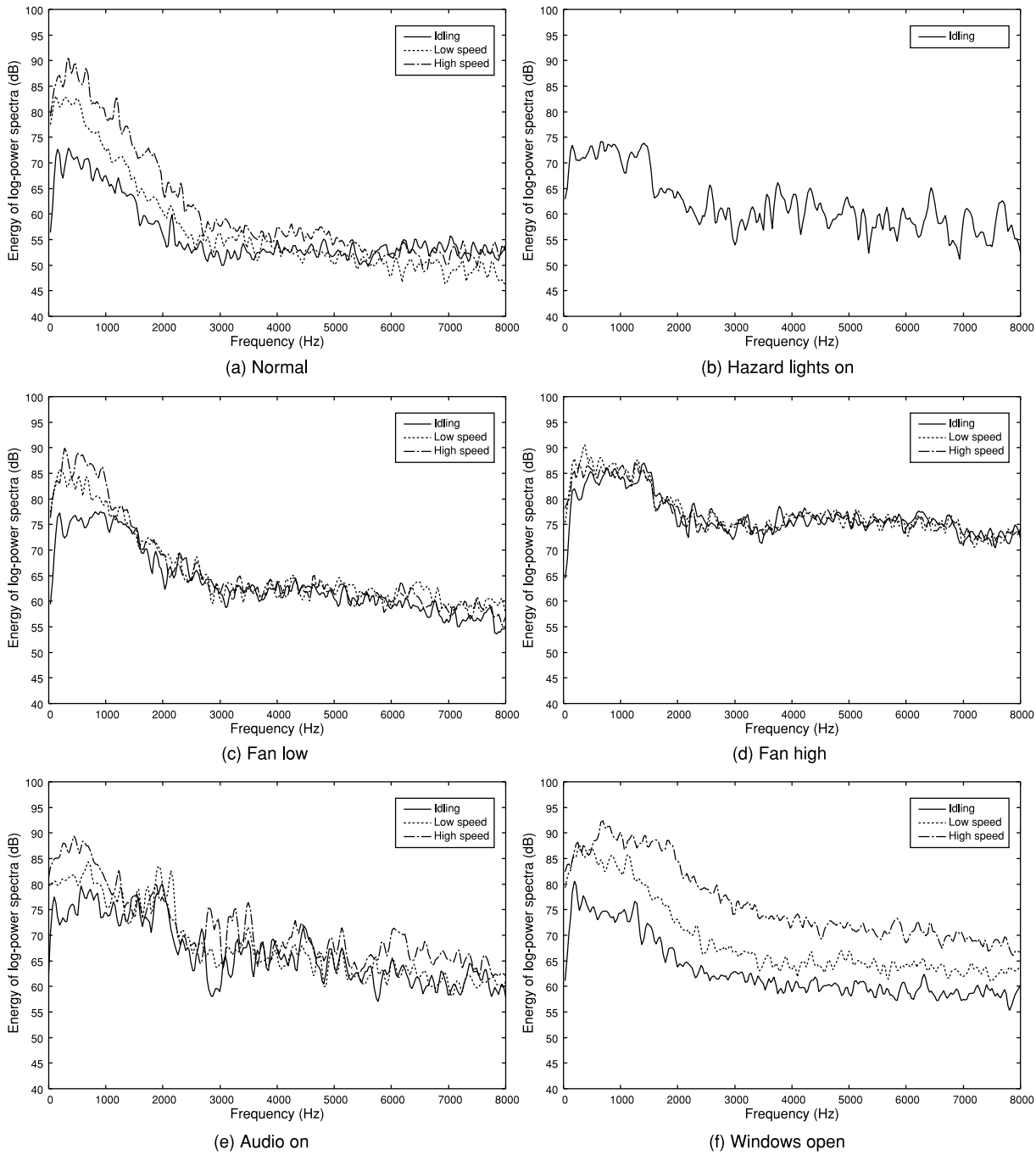


Fig. 3 Examples of average log-power spectra of adverse noises recorded in each environment with the hands-free microphone. The non-speech sections within the 50 word utterances spoken by one speaker were used for estimation of these examples. The amount of the non-speech data was approximately one minute (6,000 frames). The non-speech sections were detected by using the time labels which were given by forced alignment.

and microphone used. Speech data marked by a circle include, therefore, data on all in-car conditions (six types for idling and five types for low- or high-speed driving) spoken by all speakers (293 speakers for training and 18 speakers for testing).

For each of conditions 1, 2, and 3, data collected by using the same microphones in the same recording environment were prepared both for training and testing. These conditions correspond to the "Well-matched condition" of the AURORA3 framework [4]–[7]. Condition 4 corresponds to the "Moderately-mismatched condition" of the AURORA3 framework, of which training and test data were recorded under different conditions; that is, training and test data were collected while idling and driving, by using the same micro-

Table 4 Training data for each recording condition.

	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Condition 6	
Microphone	CT	HF	CT	HF	CT	HF	CT	HF	CT	HF	CT	HF
Idling (quiet)	○	○	○	—	—	○	—	○	○	—	○	—
Low speed	○	○	○	—	—	○	—	—	○	—	—	—

Table 5 Test data for each evaluation condition.

	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Condition 6	
Microphone	CT	HF	CT	HF	CT	HF	CT	HF	CT	HF	CT	HF
Idling	○	○	○	—	—	○	—	—	—	—	—	—
Low speed	○	○	○	—	—	○	—	○	—	○	—	○
High speed	○	○	○	—	—	○	—	○	—	○	—	○

Table 6 The number of utterances for training of each evaluation condition.

Car speed	Mic.	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
Idling (quiet)	CT	3,608	3,608	—	—	3,608	3,608
	HF	3,608	—	3,608	3,608	—	—
	Total	**7,216**	**3,608**	**3,608**	**3,608**	**3,608**	**3,608**
Low speed	CT	10,442	10,442	—	—	10,442	—
	HF	10,442	—	10,442	—	—	—
	Total	**20,884**	**10,442**	**10,442**	—	**10,442**	—
Total		**28,100**	**14,050**	**14,050**	**3,608**	**14,050**	**3,608**

phones. Both conditions 5 and 6 correspond to the "High-mismatched condition" of the AURORA3 framework, of which data collected by using different microphones under different recording conditions are used for training and testing. Tables 6 and 7 show the number of utterances for training and testing in each condition.

4. Baseline Performance

4.1 Baseline Scripts for Evaluation

The baseline scripts were designed to facilitate HMM training and evaluation by HTK [12]. The evaluation framework was designed as follows:

- All scripts are written in Perl, and work with Perl version 5 and later.
- The CENSREC-3 database provides parallel processing by multiple computers to reduce the processing time. Parallel processing is easily available by simply adding the remote host names to the configuration file of the baseline scripts.

Feature extraction

- The tool `HCopy` is used for feature extraction.
- The feature vector consisted of 12 MFCCs and log-energy with their corresponding delta and acceleration coefficients. Analysis conditions were pre-emphasis $1 - 0.97z^{-1}$, hamming window, 20-msec frame length, and 10-msec frame shift. Regarding the baseline performance, cepstral mean subtraction was not applied to the feature vectors.
- In the Mel-filter bank analysis, a cut-off was applied to frequency components lower than 250 Hz.

Acoustic model training

- The acoustic models used for speech recognition consist of triphone HMMs with five states. In HMMs trained by HTK, the initial (first) and the final (fifth) states have no distributions. Substantially, the states that do have distributions are restricted to the three center states (second to fourth states). Each distribution is represented with 32 mixture Gaussians, and there are 2,000 states that have the distributions. The topology and the number of HMM parameters are decided with respect to the standard triphone HMMs for in-car environments included in the CSRC (Continuous Speech Recognition Consortium) products [18].
- In CENSREC-3, flat-start training [19], a well known acoustic model training method, is used for estimating the HMMs' parameters. At first, the global model, i.e., the model with the global speech mean vector and diagonal variance matrix, is estimated using tool `HCompV`. Next, the initial monophone HMMs, which have parameters equal to the global ones, are constructed. The parameters of HMMs are re-estimated by iterative embedded training with the Baum-Welch estimator, tool `HERest`. The HMMs are first trained as the monophone HMMs with a single Gaussian distribution using the monophone labels. After the monophone HMM training (ten iterations of embedded training), the HMMs are converted to tied-state triphone HMMs by using tool `HHEd` and a decision tree. The iterative embedded training is also applied to the triphone HMMs by using the triphone labels. At every tenth iteration of triphone HMM training, the number of mixture Gaussians is increased by the power of two with tool `HHEd`.

Speech recognition

- Speech recognition is carried out with a Viterbi de-

Table 7 The number of utterances for testing of each evaluation condition.

Car speed	Mic.	In-car condition	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
Idling	CT	Normal	898	898	—	—	—	—
		Hazard lights on	900	900	—	—	—	—
		Fan (low)	887	887	—	—	—	—
		Fan (high)	900	900	—	—	—	—
		Audio on	896	896	—	—	—	—
		Windows open	899	899	—	—	—	—
		Total	5,380	5,380	—	—	—	—
	HF	Normal	898	—	898	—	—	—
		Hazard lights on	900	—	900	—	—	—
		Fan (low)	887	—	887	—	—	—
		Fan (high)	900	—	900	—	—	—
		Audio on	896	—	896	—	—	—
		Windows open	899	—	899	—	—	—
		Total	5,380	—	5,380	—	—	—
	Total		**10,760**	**5,380**	**5,380**	—	—	—
Low speed	CT	Normal	848	848	—	—	—	—
		Fan (low)	850	850	—	—	—	—
		Fan (high)	895	895	—	—	—	—
		Audio on	849	849	—	—	—	—
		Windows open	897	897	—	—	—	—
		Total	4,339	4,339	—	—	—	—
	HF	Normal	848	—	848	848	848	848
		Fan (low)	850	—	850	850	850	850
		Fan (high)	895	—	895	895	895	895
		Audio on	849	—	849	849	849	849
		Windows open	897	—	897	897	897	897
		Total	4,339	—	4,339	4,339	4,339	4,339
	Total		**8,678**	**4,339**	**4,339**	**4,339**	**4,339**	**4,339**
High speed	CT	Normal	900	900	—	—	—	—
		Fan (low)	900	900	—	—	—	—
		Fan (high)	900	900	—	—	—	—
		Audio on	899	899	—	—	—	—
		Windows open	898	898	—	—	—	—
		Total	4,497	4,497	—	—	—	—
	HF	Normal	900	—	900	900	900	900
		Fan (low)	900	—	900	900	900	900
		Fan (high)	900	—	900	900	900	900
		Audio on	899	—	899	899	899	899
		Windows open	898	—	898	898	898	898
		Total	4,497	—	4,497	4,497	4,497	4,497
	Total		**8,994**	**4,497**	**4,497**	**4,497**	**4,497**	**4,497**
Total			**28,432**	**14,216**	**14,216**	**8,836**	**8,836**	**8,836**

coder, tool `HVite`. As the decoding parameter, the pruning beam parameter is set to 0.0, which means that the beam search is disabled, i.e., a full search is used for decoding. The grammar scale factor is set to 0.0; i.e., the decoding is done using only acoustic scores.

- In the recognition, a standard pronunciation dictionary and recognition grammar are defined as described by the EBNF syntax notation [19] shown in Fig. 4.
- In the case of a word with connected vowels that can be pronounced by a long vowel, pronunciation rules for both the connected vowels and the long vowel are registered in the pronunciation dictionary. For example, in the case of the Japanese word "Ninshou," two pronunciation rules, "n i N sh **o u**" and "n i N sh **o:**," are registered.

In the above descriptions, tools `H*` are HTK program commands.

4.2 Baseline Recognition Results and Performance Comparison

Table 8 shows the details of baseline recognition results for each car environment for evaluation conditions 1 to 6[†] . In the table, we can see that results of matched conditions, i.e., conditions 1 to 3, are quite good. However, the word accuracy of condition 3 is lower than those of conditions 1 and 2. This is caused by low SNR of speech data recorded

[†]There may be cases where the parameters of acoustic models change slightly according to the number of computers and the operating system used for experiments. This often affects the recognition results (its fluctuation is approximately ±1%). The experiments for obtaining the baseline results were performed by using four computers with Red Hat Linux release 7.2. This phenomenon is repeatable. Hence, when you carry out the baseline evaluation with four computers, it is possible to obtain the same results as shown in Table 8.

Table 8 Details of CENSREC-3 baseline evaluation results (%).

Car speed	Mic.	In-car condition	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
Idling	CT	Normal	99.89	100.00	—	—	—	—
		Hazard lights on	99.33	99.89	—	—	—	—
		Fan (low)	99.55	100.00	—	—	—	—
		Fan (high)	97.78	99.44	—	—	—	—
		Audio on	98.77	99.67	—	—	—	—
		Windows open	99.11	99.33	—	—	—	—
		Overall	99.07	99.72	—	—	—	—
	HF	Normal	99.44	—	99.78	—	—	—
		Hazard lights on	98.78	—	98.89	—	—	—
		Fan (low)	90.19	—	94.02	—	—	—
		Fan (high)	53.56	—	53.44	—	—	—
		Audio on	81.47	—	81.36	—	—	—
		Windows open	89.66	—	89.88	—	—	—
		Overall	85.50	—	86.21	—	—	—
	Overall		**92.29**	**99.72**	**86.21**	—	—	—
Low speed	CT	Normal	100.00	100.00	—	—	—	—
		Fan (low)	100.00	100.00	—	—	—	—
		Fan (high)	97.99	98.77	—	—	—	—
		Audio on	98.82	99.41	—	—	—	—
		Windows open	99.11	98.55	—	—	—	—
		Overall	99.17	99.33	—	—	—	—
	HF	Normal	98.00	—	99.17	88.21	56.60	45.99
		Fan (low)	90.82	—	94.12	77.41	54.35	35.18
		Fan (high)	62.57	—	60.11	41.79	43.46	28.83
		Audio on	79.27	—	78.56	65.02	47.47	37.57
		Windows open	64.66	—	65.33	45.60	23.97	15.27
		Overall	78.73	—	79.10	63.17	44.92	32.33
	Overall		**88.95**	**99.33**	**79.10**	**63.17**	**44.92**	**32.33**
High speed	CT	Normal	99.89	99.89	—	—	—	—
		Fan (low)	99.67	99.89	—	—	—	—
		Fan (high)	97.67	99.22	—	—	—	—
		Audio on	99.78	99.78	—	—	—	—
		Windows open	96.66	95.21	—	—	—	—
		Overall	98.53	98.80	—	—	—	—
	HF	Normal	92.33	—	95.56	64.78	29.67	21.78
		Fan (low)	85.11	—	89.44	48.22	30.67	19.89
		Fan (high)	59.67	—	55.22	37.33	40.78	22.44
		Audio on	78.31	—	79.20	49.72	30.03	23.92
		Windows open	24.83	—	21.83	15.37	7.80	6.46
		Overall	68.07	—	68.27	43.10	27.80	18.90
	Overall		**83.30**	**98.80**	**68.27**	**43.10**	**27.80**	**18.90**
Overall			**88.43**	**99.31**	**78.36**	**52.95**	**36.20**	**25.50**

```
$words = digital_locker |
         ninsho_kaishi |
         ... |
         service_syuryo;

( [silB] $words [silE] )
```

Fig. 4 Grammar written in EBNF.

by the hands-free microphone. On the other hand, the results of mismatched conditions, i.e., conditions 4 to 6, deteriorate according to the mismatching level. Especially, the results of conditions 5 and 6, which have driving-speed and microphone mismatches, are much worse than those of the matched conditions. The most crucial in-car environment is "Windows open" one; in particular, the results of the "Windows open" environment together with high-speed driving are extremely low. In the hands-free condition, the air-conditioner environments, i.e., "Fan low" or "Fan high," also seriously degrade word accuracy. Therefore, performance improvement in these crucial environments is an important factor for in-car speech recognition, especially for the hands-free-based approach.

We have also distributed a Microsoft Excel spreadsheet to simplify the recognition performance comparison. All of the baseline results and the averaged recognition result are shown at the top of Table 9. The data entry for results (word accuracy) should be made in the middle part of Table 9, after which the relative improvement against the baseline result is automatically given in the bottom part. In Table 9, the relative improvement is calculated with the following equation:

$$\text{Relative improvement} = \frac{\%Acc - \%Acc \text{ of baseline}}{100 - \%Acc \text{ of baseline}} \times 100(\%), \quad (1)$$

where $\%Acc$ denotes the word accuracy.

Table 9 CENSREC-3 spreadsheet and the evaluation results by ETSI ES 202 050 front-end.

CENSREC-3 Evaluation Results						
CENSREC-3 Baseline Results (%)						
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Average
88.43	99.31	78.36	52.95	36.20	25.50	63.46
CENSREC-3 Word Accuracy (%)						
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Average
95.48	99.62	91.95	86.63	83.70	73.85	88.54
CENSREC-3 Relative Improvement						
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Average
60.93%	44.93%	62.80%	71.58%	74.45%	64.90%	68.63%

4.3 Evaluation Results by ETSI ES 202 050 Front-End

Table 9 also shows the evaluation results of ETSI ES 202 050 front-end.

ETSI ES 202 050 was developed as the standard front-end for distributed speech recognition that includes noise reduction, parameter compression, and so on [16]. Noise reduction of the front-end is based on Wiener filter theory and is performed in two stages as shown in Fig. 5.

The linear spectrum of each frame is estimated in the first stage, then frequency domain Wiener filter coefficients are computed by using the current frame spectrum and the spectra of the noise frames detected with VAD (Voice Activity Detection). In the PSD (Power Spectral Density) Mean block, the spectrum is smoothed along the frame index. Finally, the input signal is de-noised by filtering the time domain Mel-warped Wiener filter, which is converted from the Wiener filter for the linear frequency domain. In the second stage, additional and dynamic noise reduction is performed according to the signal-to-noise ratio of the output signal in the first stage. After this, the DC offset of the de-noised signal is removed.

The detailed results of ETSI ES 202 050 front-end are given in Table 10. This table shows that the results by ETSI ES 202 050 front-end are considerably higher than those for the baseline performance. Especially, the word accuracies of the crucial conditions, i.e., "Fan low," "Fan high," and "Windows open," are significantly improved. However, the word accuracies of the crucial environments recorded by the hands-free microphone with during high-speed driving, are not sufficient for practical use in in-car speech recognition. Thus, a continuous investigation of noise robust speech recognition is necessary for real-world applications.

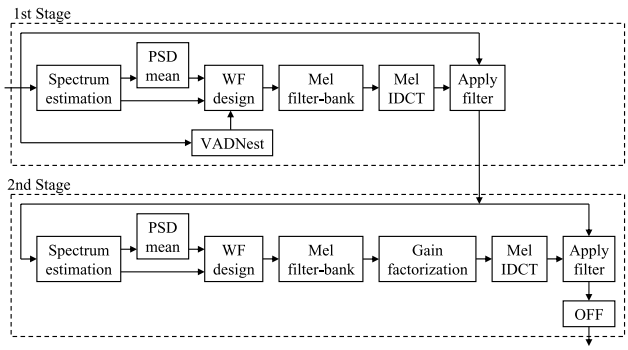


Fig. 5 Block diagram of noise reduction implemented in ETSI ES 202 050 front-end.

5. Evaluation Categories

Evaluation categories are designed for CENSREC-3 that show how much the user's method modified the baseline back-end scripts from the viewpoint of changes in the training method of HMMs, model topology, feature parameters, and so on. Users are requested to declare the category to which they belong from the following categories, according to the degree of modification to the back-end scripts from the original baseline. No changes to the back-end scripts, i.e., changes to only front-end processing, can be included in category 0. Recognition results can be fairly compared with other methods only within the same category. In addition, the following categories are borrowed from AURORA-2J with some changes.

Category 0. No changes to the back-end scripts.

Category 1. If the HMM topology is the same as the baseline scripts, any training process will be allowed. Discriminative training can be introduced in this category. The computational cost in the recognition phase should be the same as it was. Other experimental conditions are the same as in the back-end scripts.

Category 2. If the HMM topology is the same, adaptation processes can be introduced using some testing data. Speaker or environment adaptation, and PMC (Parallel Model Combination [20]) with one state noise model can be allowed in this category. An increase in the computational cost will be caused only by the adaptation process. Other experimental conditions are the same as in the back-end scripts.

Category 3. Changes in the standard HMM topology. A different number of mixtures and states can be allowed. However, the recognition unit should be the same as in the original back-end scripts ("triphone HMMs" in CENSREC-3). PMC with more than one state noise model can be included in this category. Other exper-

Table 10 Details of CENSREC-3 evaluation results by ETSI ES 202 050 front-end (%).

Car speed	Mic.	In-car condition	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
Idling	CT	Normal	99.89	99.89	—	—	—	—
		Hazard lights on	99.44	99.67	—	—	—	—
		Fan (low)	99.89	99.89	—	—	—	—
		Fan (high)	99.44	99.78	—	—	—	—
		Audio on	99.44	99.67	—	—	—	—
		Windows open	99.89	99.78	—	—	—	—
		Overall	99.67	99.78	—	—	—	—
	HF	Normal	99.78	—	100.00	—	—	—
		Hazard lights on	98.11	—	98.22	—	—	—
		Fan (low)	98.53	—	98.65	—	—	—
		Fan (high)	84.33	—	85.11	—	—	—
		Audio on	89.62	—	90.07	—	—	—
		Windows open	96.77	—	97.22	—	—	—
		Overall	94.52	—	94.87	—	—	—
	Overall		**97.09**	**99.78**	**94.87**	—	—	—
Low speed	CT	Normal	99.88	99.76	—	—	—	—
		Fan (low)	100.00	99.88	—	—	—	—
		Fan (high)	99.22	99.33	—	—	—	—
		Audio on	99.18	99.53	—	—	—	—
		Windows open	99.55	99.11	—	—	—	—
		Overall	99.56	99.52	—	—	—	—
	HF	Normal	98.82	—	99.17	97.52	96.11	91.51
		Fan (low)	97.76	—	97.53	96.71	94.94	89.41
		Fan (high)	87.37	—	89.05	84.92	80.34	66.82
		Audio on	91.05	—	91.05	84.81	87.28	75.85
		Windows open	86.85	—	86.85	83.50	78.82	68.23
		Overall	92.26	—	92.63	89.38	87.32	78.13
	Overall		**95.91**	**99.52**	**92.63**	**89.38**	**87.32**	**78.13**
High speed	CT	Normal	100.00	99.78	—	—	—	—
		Fan (low)	100.00	99.89	—	—	—	—
		Fan (high)	99.33	99.33	—	—	—	—
		Audio on	98.89	99.44	—	—	—	—
		Windows open	99.22	99.22	—	—	—	—
		Overall	99.49	99.53	—	—	—	—
	HF	Normal	97.00	—	98.67	95.78	91.89	84.67
		Fan (low)	95.33	—	96.22	93.89	89.33	83.11
		Fan (high)	85.22	—	86.00	83.00	78.89	64.00
		Audio on	92.10	—	92.55	87.43	86.54	75.42
		Windows open	64.14	—	65.48	59.80	54.34	41.31
		Overall	86.77	—	87.79	83.99	80.21	69.71
	Overall		**93.13**	**99.53**	**87.79**	**83.99**	**80.21**	**69.71**
Overall			**95.48**	**99.62**	**91.95**	**86.63**	**83.70**	**73.85**

imental conditions are the same as in the back-end scripts.

Category 4. Any process will be allowed as long as the decoder is the same as in the original back-end scripts (`HVite` in CENSREC-3). Changes to a model unit, syntax and lexicon for the decoder can be included in this category.

Category 5. Any process with any computational cost will be allowed.

Category B. The use of any training data not included in CENSREC-3 — not only speech data, but also environment noise data. Of course, CENSREC-3 constitutes the evaluation data. This category essentially differs from categories 1 to 5.

6. Conclusions

In this paper, we introduced CENSREC-3, an evaluation framework for Japanese in-car speech recognition, and presented the evaluation results by ETSI ES 202 050 front-end. We also indicated the crucial conditions for practical use of the in-car speech recognition.

In the near future, we will gradually design and distribute the evaluation frameworks of noisy speech recognition for increasingly difficult conditions; i.e., non-stationary noise environments, reverberant environments, large-vocabulary continuous-speech recognition tasks, and so on. We also plan to develop and distribute a noise database for noisy speech recognition, alternative evaluation measures to word accuracy. and a tool kit of conventionally used noise compensation methods. The latest information about CENSREC will be provided on the following Website.

CENSREC Website:
http://sp.shinshu-u.ac.jp/CENSREC/

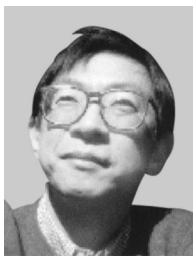
Acknowledgements

We would like to thank the members of the IPSJ SIG-SLP Noisy Speech Recognition Evaluation Working Group.

References

[1] DARPA project Website, http://www.nist.gov/speech/publications/
[2] SPINE Website, http://elazar.itd.nrl.navy.mil/spine/
[3] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," Proc. ISCA ITRW ASR2000, pp.18–20, Paris, France, Sept. 2000.
[4] AU/378/01, "Danish speechdat-car digits database for ETSI STQ-Aurora advanced DSR," Aalborg University, Jan. 2001.
[5] AU/225/00, "Baseline results for subset of speechdat-Car finnish database for ETSI STQ WI008 advanced front-end evaluation," Nokia, Jan. 2000
[6] AU/273/00, "Description and baseline results for the subset of the speechdat-car German database used for ETSI STQ Aurora WI008 advanced DSR front-end evaluation," Texas Instruments, Dec. 2001.
[7] AU/271/00, "Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: Description and baseline results," UPC, Nov. 2000.
[8] AU/337/01, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task: Version 1.0," Ericsson, June 2001.
[9] AU/345/01, "Large vocabulary evaluation of front-ends: Baseline recognition system description, final report," Mississippi State University, Jan. 2002.
[10] ETSI Website, http://www.etsi.org/
[11] R.G. Leonard, "A database for speaker independent digit recognition," Proc. ICASSP '84, vol.3, pp.328–331, San Diego, USA, March 1984.
[12] HTK Website, http://htk.eng.cam.ac.uk/
[13] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & What next," Proc. Eurospeech '01, Aalborg, Denmark, Sept. 2001.
[14] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.535–544, March 2005.
[15] M. Fujimoto, K. Takeda, and S. Nakamura, "CENSREC-2: Data collection for in-car digit speech recognition and its common evaluation framework," IPSJ SIG Technical Reports, SLP-60-3, pp.13–18, Feb. 2006.
[16] ETSI ES 202 050 v.1.1.4, "Speech processing, transmission and quality aspects (STQ), advanced distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," Nov. 2005.
[17] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, and F. Itakura, "Construction and evaluation of a large in-car speech corpus," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.553–561, March 2005.
[18] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yadama, A, Ito, K. Itou, and K. Shikano, "Continuous speech recognition consortium — An open repository for CSR tools and models," Proc. LREC '02, pp.1438–1441, Las Palmas, Spain, May 2002.
[19] S. Young, "The HTK book," Entropic Cambridge Research Laboratory Ltd., 1999.
[20] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," IEEE Trans. Speech Audio Process., vol.4, no.5, pp.352–359, May 1996.

Masakiyo Fujimoto received B.E., M.E., and Doctor of Engineering degrees all from Ryukoku University in 1997, 2001, and 2005, respectively. From 2004–2006, he worked with ATR Spoken Language Communication Research Laboratories. He is currently a research associate at NTT Communication Science Laboratories. He received the Awaya Award from Acoustical Society of Japan (ASJ) in 2003. His current research interests include noise-robust speech recognition. He is a member of the ASJ and IEEE.



Kazuya Takeda received his B.E., M.E., and Doctor of Engineering degrees all from Nagoya University in 1983, 1985 and 1994, respectively. In 1986, he joined ATR, where he involved in two major projects on speech database construction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated in a project to construct a voice-activated telephone extension system. Since 1995, he has been working for Nagoya University. He is a professor and a leader of the acoustic, speech and behavioral signal processing group at the Graduate School of Information Science, Nagoya University.

Satoshi Nakamura received his B.S. degree in electronic engineering from Kyoto Institute of Technology in 1981 and a Ph.D. degree in information science from Kyoto University in 1992. Between 1981–1993, he worked with the Central Research Laboratory, Sharp Corporation, Nara, Japan. From 1986–1989, he worked with ATR Interpreting Telephony Research Laboratories. From 1994–2000, he worked as an associate professor of the Graduate School of Information science at Nara Institute of Science and Technology, Japan. In 1996, he was a visiting research professor of the CAIP center of Rutgers University of New Jersey USA. In 2000 he moved to ATR spoken language translation research labs to serve as the department head for acoustics and speech research. He is currently the director at ATR Spoken Language Communication Laboratories, Japan and a group leader of Spoken Language Communication Group at National Institute of Information and Communications Technology. He also serves as an honorary professor at the University of Karlsruhe, Germany since 2004. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He is a member of the Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ), and IEEE. He received the Awaya award from the Acoustical Society of Japan in 1992, and the Interaction2001 best paper award from the Information Processing Society of Japan in 2001.